

Scalable Bayes under Informative Sampling

Terrance D. Savitsky*, Sanvesh Srivastava†

October 27, 2016

Abstract

The United States Bureau of Labor Statistics collects data using survey instruments under informative sampling designs that assign probabilities of inclusion to be correlated with the response. The bureau extensively uses Bayesian hierarchical models and posterior sampling to impute missing items in respondent-level data and to infer population parameters. Posterior sampling for survey data collected based on informative designs are computationally expensive and do not support production schedules of the bureau. Motivated by this problem, we propose a new method to scale Bayesian computations in informative sampling designs. Our method divides the data into smaller subsets, performs posterior sampling in parallel for every subset, and combines the collection of posterior samples from all the subsets through their mean in the Wasserstein space of order 2. Theoretically, we construct conditions on a class of sampling designs where posterior consistency of the proposed method is achieved. Empirically, we demonstrate that our method is competitive with traditional methods while being significantly faster in many simulations and in the Current Employment Statistics survey conducted by the bureau.

*U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E, Washington, D.C. 20212 USA

†Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, USA

Key words: Barycenter; Distributed Bayesian computations; Pseudo posterior distribution; Survey sampling; Wasserstein distance; Posterior consistency; Monte Carlo; Current Employment Statistics survey.

1 Introduction

Bayesian hierarchical models are popular for inference and imputation in complex data because latent dispositional states that underlie observed behaviors can be directly parameterized (Savitsky & Dalal 2013). These models are extensively employed for estimations on data acquired from surveys. Survey data are often collected using sampling schemes which imply that inclusion probabilities for units are correlated with the response variable of interest. Bayesian inference remains the gold standard for imputation of missing data and for capturing uncertainty in the estimation of parameters but requires tuning under informative sampling (Gelman et al. 2013). There are many Bayesian methods that account for informative sampling designs, but they focus on summary statistics. In contrast, few Bayesian methods are available for imputation of missing responses and for inference on survey units (Savitsky & Toth 2016). Often, however, these methods cannot be applied to informative survey sampling data because posterior computations become intractable due to the size of the data. Motivated by this problem, we propose a new method based on the divide-and-conquer technique that allows existing Bayesian methods for informative sampling to be scaled to modern survey sampling data with minimal modifications.

Existing Bayesian methods that account for sampling informativeness focus on inference on summary statistics. These methods employ an empirical likelihood that incorporates sampling weights. The sampling weight of a unit is inversely proportional to its first order marginal inclusion probability. This balances the information in the observed sample to approximate information in the finite population (Dong et al. 2014, Kuniyama et al. 2016, Rao & Wu 2010,

Si et al. 2015). A major limitation of these methods is that inference is restricted to summary statistics; therefore, these methods cannot be used for imputation of missing responses in any survey data or for inference on the parameters of the model.

Our focus in this work is two fold. First, we construct a general formulation for the posterior distribution of model parameters in informative sampling when the focus is inference on functions of parameters and imputation, rather than solely design-based inference. This is useful for survey data where the interest lies in inference about the parameters that are part of the hierarchical model. Key to our approach is the sampling-weighted pseudo posterior distribution that corrects for an informative sampling design and that permits inference from any model specified by the data analyst (Savitsky & Toth 2016). This approach constructs a pseudo posterior density by the exponentiation of each unit likelihood by its sampling weight to provide a noisy approximation of the population posterior density. Theoretical results guarantee frequentist L_1 contraction of the sampling-weighted pseudo posterior distribution to the true generating model. Despite the attractive theoretical guarantees, performing pseudo posterior inference is computationally expensive on the large-sized and respondent level response, including households or establishments, typically collected by the United States Federal Statistical agencies. This motivates our second contribution in which we extend and generalize the *Wasserstein Posterior* approach for scalable Bayesian inference due to Srivastava et al. (2015) to account for informative sampling designs. This extension is extremely efficient and supports the rapid turnaround cycles used by the bureau to publish the employment statistics on a monthly basis.

The models and scalable estimation procedures for data collected under informative sampling designs that we develop in this work are applicable to the typically-used sampling designs. Our method consists of three steps. First, the sampled units are randomly split into disjoint subset such that computations for posterior sampling in any subset is tractable. Second, we modify the sampling scheme for every subset to account for the informativeness in the

design and to ensure that the variances of the posterior distributions conditioned on any data subset and on the full data are of the same order. Third, we use empirical measures supported on subset posterior samples to approximate the different subset posteriors and combine these empirical measures through their barycenter in their Wasserstein space of order 2. Our method generalizes the *Wasserstein Posterior* of Srivastava et al. (2015) in the second step where we account for informativeness of the design. The computation of a barycenter from subset posteriors scales sublinearly in sample size because each subset posterior estimation may be run in parallel, limited only by computational resources. The proposed method is applicable to a variety of sampling schemes based on informative designs. Our theoretical results show that if the number of subsets are chosen appropriately, then the Wasserstein Posterior under informative sampling converges to the true parameter at a near optimal rate.

2 Motivating Data: The Current Employment Statistics Survey

The United States Bureau of Labor Statistics administers the Current Employment Statistics survey to non-farm, public, and private business establishments across the United States on a monthly basis, receiving approximately 270,000 submitted responses in each month. The estimated total employment is published for detailed industry categories by state and for selected metropolitan areas. The survey uses a stratified sampling design with strata constructed by combinations of state, broad industry grouping, and employment size divided into 8 categories. The business establishments are sampled by their unique unemployment insurance tax identification numbers, which may contain a cluster of multiple individual sites. If a business establishment is selected based on its unique identification number, then all of the associated sites in that cluster are also included. Stratum-indexed inclusion probabilities are set to be

proportional to the average employment size for member establishments of that stratum.

The Current Employment Statistics survey constructs a *known* sampling design distribution that assigns higher inclusion probabilities to establishments with a relatively larger number of employees. This is a proportion-to-size design that induces a correlation among sample inclusion probabilities and total employment; larger establishments more strongly influence the variance of domain-indexed total employment statistics published by the bureau. Such sampling designs are called *informative* because they induce a correlation between selection probabilities and observed values. In this survey, distributions of establishment employment counts for samples will be skewed to higher values than present in the underlying population. If the informativeness in the design is not modeled, then inference on population parameters conditional on the survey data will be biased (Savitsky & Toth 2016).

There is a short time gap between the receipt of establishment submissions at the end of a month and the subsequent publication of employment estimates for that month; the imputation for missing items and estimation of employment statistics must be performed quickly. The relatively large number of submissions with non-zero changes in employment levels, coupled with the rapid publication schedule, require the use of computationally scalable estimation tools. The sampling-weighted pseudo posterior distribution proposed in Savitsky & Toth (2016) fails to meet these requirements, motivating our method based on the Wasserstein Posterior for computationally efficient imputation of missing responses and inferences on model parameters.

3 Generalizing Stochastic Approximation

3.1 Preliminaries: Wasserstein Barycenter

The order 2 Wasserstein space probability measures, denoted as $\mathcal{P}_2(\Theta)$, on a separable and complete metric space, denoted as (Θ, ρ) , is

$$\mathcal{P}_2(\Theta) = \left\{ \mu : \int_{\theta \in \Theta} \rho(\theta_0, \theta)^2 \mu(d\theta) < \infty \right\}, \quad (1)$$

where $\mathcal{P}_2(\Theta)$ does not depend on choice of θ_0 . The companion order 2 Wasserstein distance for measures, $\mu, \omega \in \mathcal{P}_2(\Theta)$, where $\Pi(\mu, \omega)$ is the set of product probability measures on $\Theta \times \Theta$ with marginals, μ and ω , is defined to be,

$$W_2(\mu, \omega) = \left(\inf_{\pi \in \Pi(\mu, \omega)} \int_{\Theta \times \Theta} \rho^2(x, y) d\pi(x, y) \right)^{\frac{1}{2}}.$$

Let Π_1, \dots, Π_K be K probability measures in $\mathcal{P}_2(\Theta)$, then Agueh & Carlier (2011) define the barycenter of Π_1, \dots, Π_K as

$$\bar{\Pi} = \operatorname{argmin}_{\Pi \in \mathcal{P}_2(\Theta)} \frac{1}{K} \sum_{j=1}^K W_2^2(\Pi, \Pi_j), \quad (2)$$

where $\bar{\Pi}$ exists uniquely and belongs to the set $\mathcal{P}_2(\Theta)$.

The Wasserstein barycenter motivates the *Wasserstein Posterior* approach for scalable Bayesian inference (Srivastava et al. 2015). Let y_1, \dots, y_N be the finite population data. Suppose we divide the data into K equally-sized subsets of size M such that $N = KM$ and subset j includes data $y_{[j]} = \{y_{j1}, \dots, y_{jM}\}$ ($j = 1, \dots, K$). Suppose $\Pi_M(\cdot \mid y_{[j]})$ and $\Pi_N(\cdot \mid y_1, \dots, y_N)$ are posterior distributions for $\theta \in \Theta$ conditioned on subset j and full data, respectively, such that the variances of $\Pi_M(\cdot \mid y_{[j]})$ and $\Pi_N(\cdot \mid y_1, \dots, y_N)$ are of the same order. The *Wasserstein Posterior*, denoted as $\bar{\Pi}(\cdot \mid y_1, \dots, y_N)$, is the Wasserstein barycenter of $\Pi_M(\cdot \mid y_{[j]})$ ($j = 1, \dots, K$) defined in (2). If posterior samples are available from $\Pi_j(\cdot \mid y_{[j]})$ ($j = 1, \dots, K$), then an empirical approximation of $\bar{\Pi}(\cdot \mid y_1, \dots, y_N)$ can be estimated by solving a linear program; see Srivastava et al. (2015) for details.

3.2 Generalized Stochastic Approximation

Suppose there exists a Lebesgue measurable population-generating density, $\pi(y_i|\theta)$, for unit (e.g., establishment) $i \in U$, where U denotes a finite population and $|U| = N$. Without loss of generality, suppose we divide the finite population units into K disjoint subsets, $\{U_j\}$ ($j = 1, \dots, K$), of equal size, $|U_j| = M$, ($j = 1, \dots, K$). Under random sampling of the finite population, we don't observe the full population, U , but a sample $S \subset U$ and $|S| = n \leq N$, where $|S|$ represents the number of elements in S . Let $\delta_i \in \{0, 1\}$ denote the sample inclusion indicator for units $i = 1, \dots, N$ from the population. The density for the observed sample is denoted by, $\pi(y_{o,1}, \dots, y_{o,n} | \theta) = \pi(\{y_i : \delta_i = 1, i = 1, \dots, N\} | \theta)$, where “o” indicates “observed.”

Savitsky & Toth (2016) define a pseudo posterior distribution tuned for the theoretical setup of informative sampling. They construct a plug-in approximation for the finite population posterior density estimated on the observed sample as

$$\pi^\pi(\theta | y_{o,1}, \dots, y_{o,n}, \tilde{w}_1, \dots, \tilde{w}_n) \propto \left\{ \prod_{i=1}^n p(y_{o,i} | \theta)^{\tilde{w}_i} \right\} \pi(\theta), \quad (3)$$

where $\pi(\theta)$ is the prior parameter density, $\tilde{w}_i = nw_i(\sum_{i=1}^n w_i)^{-1}$ ($i = 1, \dots, n$), and π_i is the marginal inclusion probability of unit i . We recover the full-data posterior density from (3) if we set $\tilde{w}_i = 1$ ($i = 1, \dots, n$). The exponent \tilde{w}_i corrects for sampling informativeness and ensures that \tilde{w}_i assigns the relative importance of the likelihood contribution of unit i to approximate the likelihood for the population. The scaling factor here is 1 in that weights are scaled to the sample size, n , which asymptotically expresses the amount of information present in our sample.

The sampled observations are often dependent in design distributions under the informative sampling. Savitsky & Toth (2016) define a condition under which the sampling design distribution produces samples which are asymptotically independent as the finite population size, N , increases, which is needed to guarantee L_1 contraction. Many sampling designs obey

this condition, in practice, including that for the Current Employment Statistics survey, where the number of establishments increases within each industry and state in the limit. There are 2 additional conditions that restrict the class of sampling designs required for consistency and they are formally reviewed in Section 4. We will drop the subscript “ o ” in y_o in the sequel because our focus is on data acquired from a sample of a finite population.

In many applications sampling from the density in (3) is computationally expensive and it is easier to sample from the posterior density conditioned on a data subset. The observed sample, S , is first divided into disjoint K disjoint subsets, S_j ($j = 1, \dots, K$), each of equal size, $m = |S_j| = n/K$, such that $S = S_1 \cup \dots \cup S_K$, where equal subset sizes are assumed for ease of presentation. We construct a pseudo likelihood for density, $p(y_{ji} | \theta)$, for unit $i \in S_j$, by exponentiating it with its sample weight, \tilde{w}_{ji} , to form,

$$\pi^\pi(\theta | y_{[j]}) \propto \left(\prod_{i=1}^m p(y_{ji} | \theta)^{\tilde{w}_{ji}} \right) \pi(\theta) \quad (4)$$

We redefine \tilde{w}_{ji} as $nw_i(\sum_{i \in S_j} w_i)^{-1}$ ($j = 1, \dots, K$) such that variance of θ with density $\pi^\pi(\theta | y_{[j]})$ ($j = 1, \dots, K$) is of the same order as that of $\pi^\pi(\theta | y_{o,1}, \dots, y_{o,n}, \tilde{w}_1, \dots, \tilde{w}_n)$ in (3). This ensures that all subset pseudo posterior distribution are noisy approximations of the full-sample pseudo posterior distribution.

Our formulation of subset pseudo posterior density in (4) generalizes the stochastic approximation in Srivastava et al. (2015). The estimation of Wasserstein Posterior using (2) requires that variances of subset and full-data pseudo posterior distributions are of the same order. Srivastava et al. (2015) suggest to raise the likelihood of every observation to the power of K . If we fix $\tilde{w}_{ji} = K$ for every i and j in (4), then we recover the subset posterior density of Srivastava et al. (2015). The differential weighting of each unit likelihood contribution is key in balancing the information in a sample to approximate the information about θ in the finite population and in subset j . The *generalized Wasserstein posterior* of the scaled subset pseudo posterior distributions in (4) is computed using (2) and provides an approximation to

the partially-observed finite population posterior distribution under informative sampling. We study next the theoretical properties of the scaled pseudo posterior distribution and the generalized Wasserstein posterior computed using K scaled subset pseudo posterior distributions.

4 Consistency of the Generalized Wasserstein Posterior

4.1 Setup

Consider the theoretical setup for an informative sampling design. Let ν be a positive integer, and U_ν is a finite population of size $|U_\nu| = N_\nu$ such that if $\nu < \nu'$, then $N_\nu < N_{\nu'}$. The sequence $|U_\nu|$ increases to ∞ as ν increases to ∞ . Let $Y_{\nu 1}, \dots, Y_{\nu N_\nu}$ be a sequence of independent and non-identically distributed random variables that are defined on population U_ν and take values on the measurable product space $\otimes_{i=1}^{N_\nu} (\mathcal{Y}_{\nu i}, \mathcal{A}_{\nu i})$, where $\mathcal{A}_{\nu i}$ is the Borel sigma-algebra on $\mathcal{Y}_{\nu i}$ ($i = 1, \dots, N_\nu$). For any parameter $\theta \in \Theta \subset \mathbb{R}^p$, let $P_{\theta \nu i}$ represent the probability distribution of $Y_{\nu i}$ indexed by θ that has the density $dP_{\theta \nu i} = p(y_{\nu i} | \theta) d\theta$ relative to a sigma-finite measure μ_i ($i = 1, \dots, N_\nu$). Define the product measure $P_\theta^{N_\nu}$ on $\otimes_{i=1}^{N_\nu} (\mathcal{Y}_{\nu i}, \mathcal{A}_{\nu i})$ as $P_\theta^{N_\nu} = \otimes_{i=1}^{N_\nu} P_{\theta i}$ that has density $\prod_{i=1}^{N_\nu} p(y_{\nu i} | \theta)$ with respect to $\otimes_{i=1}^{N_\nu} \mu_i$. We write $Y_{\nu i}$, $y_{\nu i}$, and $P_\theta^{N_\nu}$ as Y_i , y_i , and P_θ for brevity in the remainder of the paper because the context is clear.

4.2 Pseudo Posterior Distribution

We employ two equivalent approaches to estimate posterior distributions of parameters based on two sampling schemes. In the first or usual approach, the observed data are sampled from a finite population under a survey sampling design and the posterior distribution of parameters is obtained given the observed data. Second, the finite population is divided into K disjoint finite sub-populations followed by sampling in each sub-population. We calculate K posterior

distributions of parameters given the observed data in each sub-population, which are called the subset pseudo posterior distributions. These K posterior distributions are combined using the generalized Wasserstein posterior.

The first sampling scheme is concerned with the finite population U_ν for any ν . The observed data are sampled from the finite population, U_ν , under a survey sampling design that induces a known distribution, P_ν , defined on a vector of random inclusion indicators for the population units, $\delta_\nu = (\delta_{\nu 1}, \dots, \delta_{\nu N_\nu})$, where $\delta_{\nu i} \in \{0, 1\}$ indexes inclusion of unit i in observed sample, S_ν . The joint distribution over $(\delta_{\nu 1}, \dots, \delta_{\nu N_\nu})$ is described by known marginal unit inclusion probabilities, $\pi_{\nu i} = \text{pr}\{\delta_{\nu i} = 1\}$ for all $i \in U_\nu$ and the second-order pairwise probabilities, $\pi_{\nu i\ell} = \text{pr}\{\delta_{\nu i} = 1 \cap \delta_{\nu \ell} = 1\}$ for $i, \ell \in U_\nu$. The posterior distribution of θ given the observed data is $\Pi^\pi(\theta \mid \{y_i : \delta_{\nu i} = 1, i = 1, \dots, N_\nu\})$.

The second sampling scheme starts with the finite sub-populations obtained from U_ν . Let $U_{\nu j}$ ($j = 1, \dots, K$) be the collection of finite sub-populations, each of size M_ν , such that $U_\nu = U_{\nu 1} \cup \dots \cup U_{\nu K}$ and $N_\nu = KM_\nu$. The K populations are all generated from, P_θ , with $dP_\theta = p(y_{ji} \mid \theta) d\theta$. The resulting set of K samples are typically dependent due to the without replacement sampling design, where the inclusion probability of a unit in $U_{\nu j}$, $j \in \{1, \dots, K\}$, will depend on whether units in $U_{\nu \ell}$, $\ell \neq j \in \{1, \dots, K\}$, are co-included. The two steps of drawing a sample of observed data from the finite population and subsequent division into disjoint subsets from the first approach are re-cast as a single informative without replacement sampling step from the collection of K disjoint finite populations in the second equivalent approach. This second approach is used to construct our theoretical results that follow. We extend notations, $\pi_{\nu ji} = \text{pr}\{\delta_{\nu ji} = 1\}$ and $\pi_{\nu j\ell} = \text{pr}\{\delta_{\nu ji} = 1 \cap \delta_{\nu j\ell} = 1\}$ for $i, \ell \in U_{\nu j}$.

Our task is to perform inference about the unknown true, θ_0 , that we suppose generates the finite population from P_{θ_0} , by assigning a prior $\Pi(\theta)$ with density $\pi(\theta)$. We construct a

sampling-weighted pseudo likelihood as in Savitsky & Toth (2016) by defining

$$p_{\theta ji}^\pi = p(y_{ji} \mid \theta)^{\frac{\delta_{\nu ji}}{\pi_{\nu ji}}}, \quad i \in U_{\nu j}. \quad (5)$$

The likelihood contribution of sample i in subset j is weighted by $\pi_{\nu ji}^{-1}$ in (5) so that the information in subset j approximates the information in partially observed finite population of size M_ν . We use the pseudo likelihood in (5) and the prior $\pi(\theta)$ to obtain the pseudo posterior density for subset j as

$$\pi_j^\pi(\theta \mid y_{[j]}, \delta_{\nu[j]}) = \frac{\prod_{i \in [j]} \frac{p_{\theta ji}^\pi}{p_{\theta_0 ji}^\pi} \pi(\theta)}{\int_{\Theta} \prod_{i \in [j]} \frac{p_{\theta ji}^\pi}{p_{\theta_0 ji}^\pi} \pi(\theta) d\theta}, \quad (6)$$

where $[j] = \{i \in U_{\nu j}\}$ denotes the M_ν finite population units in $U_{\nu j}$, $y_{[j]} = \{y_{ji} : i \in U_{\nu j}\}$, and $\delta_{\nu[j]} = \{\delta_{\nu ji} : i \in U_{\nu j}\}$. The sampling weights $\pi_{\nu ji}$ ($i \in S_{\nu j}$) in the observed subsample, $S_{\nu j} \subseteq U_{\nu j}$, satisfy $\sum_{i \in S_{\nu j}} \pi_{\nu ji}^{-1} = n_\nu$ so that $\pi_j^\pi(\theta \mid y_{[j]}, \delta_{\nu[j]})$ ($j = 1, \dots, K$) is a noisy approximation of the posterior density defined on the observed sample of size n_ν , $\pi(\theta \mid \{y_i : \delta_{\nu i} = 1, i = 1, \dots, N_\nu\})$. We recover the subset pseudo posterior density defined in Srivastava et al. (2015) if we set $\delta_{\nu[j]} = (1, \dots, 1)$ in (6).

We use the generalized Wasserstein posterior to combine K subset posterior distributions estimated using (6). Let $\Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]})$ ($j = 1, \dots, K$) represent the K subset posterior posteriors and $\bar{\Pi}^\pi(\cdot \mid \{y_i : \delta_{\nu i} = 1, i = 1, \dots, N_\nu\})$ represent the generalized Wasserstein posterior. The event probabilities in the informative sampling designs are represented by P_{θ_0, P_ν} , which is indexed by θ_0 and P_ν to indicate the joint distribution with respect to generation of the finite population and subsequent taking of the observed sample. The resulting sample observations taken from $U_{\nu j}$ under P_ν are now dependent due to the dependence induced by sampling without replacement. We extend the definition of $\mathcal{P}_2(\Theta)$ in (1) to define

$$\mathcal{P}_{2\nu}(\Theta) = \left\{ \mu_\nu : \int_{\theta \in \Theta} \rho_\nu(\theta_0, \theta)^2 \mu_\nu(d\theta) < \infty \right\}.$$

Assuming $\Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]}) \in \mathcal{P}_{2\nu}(\Theta)$ ($j = 1, \dots, k$), we extend the definition in (2) to define

the generalized Wasserstein posterior as

$$\bar{\Pi}^\pi = \operatorname{argmin}_{\Pi \in \mathcal{P}_{2\nu}(\Theta)} \frac{1}{K} \sum_{j=1}^K W_2^2(\Pi, \Pi_j^\pi),$$

and Proposition 3.8 in Agueh & Carlier (2011) implies that $\bar{\Pi}^\pi$ exists uniquely in $\mathcal{P}_{2\nu}(\Theta)$.

4.3 Empirical process functionals

We will approximate the joint distribution for population generation and informative sampling using an empirical distribution construction similar to Breslow & Wellner (2007) that incorporates inverse inclusion probability weights, $1/\pi_{\nu ji}$ ($i = 1, \dots, M_\nu$),

$$P_{M_{\nu j}}^\pi = \frac{1}{M_\nu} \sum_{i=1}^{M_\nu} \frac{\delta_{\nu ji}}{\pi_{\nu ji}} \delta(Y_{ji}), \quad (7)$$

where $\delta(Y_{ji})$ denotes the Dirac delta function, with probability mass 1 on observed Y_{ji} and we recall that $M_\nu = |U_{\nu j}|$ denotes the size of the finite population for subset j .

We follow the notational convention of Ghosal et al. (2000) and define the associated expectation functionals with respect to these empirical distributions by $P_{M_{\nu j}}^\pi f = \frac{1}{M_\nu} \sum_{i=1}^{M_\nu} \frac{\delta_{\nu ji}}{\pi_{\nu ji}} f(Y_{ji})$. Similarly, $P_{M_{\nu j}} f = \frac{1}{M_\nu} \sum_{i=1}^{M_\nu} f(Y_{ji})$ for $f : \mathcal{Y} \rightarrow \mathfrak{R}$. Associated centered empirical processes are defined, $G_{M_\nu}^{\pi j} = \sqrt{M_\nu} (P_{M_{\nu j}}^\pi - P_0)$ and $G_{M_{\nu j}} = \sqrt{M_\nu} (P_{M_{\nu j}} - P_0)$.

The sampling-weighted, pseudo Hellinger distance between $P_{\theta_{j,1}}, P_{\theta_{j,2}} \in \{\otimes_{i=1}^{M_\nu} P_{\theta,ji} : \theta \in \Theta\}$, $h_{M_\nu}^{\pi,2}(\theta_1, \theta_2) = \frac{1}{M_\nu} \sum_{i=1}^{M_\nu} \frac{\delta_{\nu ji}}{\pi_{\nu ji}} h^2(p_{\theta_{1,ji}}, p_{\theta_{2,ji}})$, where $h(p_1, p_2) = \left\{ \int (\sqrt{p_1} - \sqrt{p_2})^2 d\nu \right\}^{\frac{1}{2}}$ for dominating measure, ν . The associated non-sampling Hellinger distance is specified with, $h_{M_\nu}^2(\theta_1, \theta_2) = \frac{1}{M_\nu} \sum_{i=1}^{M_\nu} h^2(p_{\theta_{1,ji}}, p_{\theta_{2,ji}})$.

4.4 Main Results

We next specify six conditions for the space, (Θ, ρ) , and the associated prior on the space, Π , followed by the three additional conditions on the sampling design distribution, P_ν . Suppose we have a sequence, $\epsilon_{M_\nu} \downarrow 0$ and $M_\nu \epsilon_{M_\nu}^2 \uparrow \infty$ and $m_\nu \epsilon_{M_\nu}^2 \uparrow \infty$ as positive integer $\nu \uparrow \infty$,

(A1) (Compactness) Θ is a compact space in the ρ metric and θ_0 is an interior point of Θ .

(A2) (Pseudo Distance bounded from below) For any $\theta_1, \theta_2 \in \Theta$ and $j = 1, \dots, K$, there exists a positive constant, C_L , such that:

$$h_{M_{\nu j}}^2(\theta_1, \theta_2) \geq C_L \rho^2(\theta_1, \theta_2).$$

(A3) (Local entropy condition - Size of model) Let constants $D_1 > 0$ and $0 < D_2 < \frac{D_1^2}{2^{12}}$, and define a function, $\Phi(u, r) \geq 0$, increasing in $u \in \mathbb{R}^+$, non-decreasing in $r \in \mathbb{R}^+$, such that for all sufficiently large M_ν ,

$$H_{[]} (u, \{\theta \in \Theta : h_{M_{\nu j}}(\theta, \theta_0) \leq r\}, h_{M_{\nu j}}) \leq \Phi(u, r),$$

where $H_{[]}$ denotes the $h_{M_{\nu j}}$ -bracketing entropy, which is the log of 1+ the bracketing number defined for data drawn independently in Srivastava et al. (arXiv:1508.05880), and the size of the bracketing entropy bound is restricted to,

$$\int_{D_1 \frac{r^2}{12}}^{D_1 r} \sqrt{\Phi(u, r)} du < D_2 \sqrt{M_{\nu j}} r^2.$$

(A4) (Prior thickness) There exist positive constants, κ and c_π such that uniformly over all $j = 1, \dots, K$,

$$\Pi \left\{ \theta \in \Theta : \frac{1}{M_\nu} \sum_{i=1}^{M_\nu} E_{P_{\theta_0}} \exp \left(\kappa \log_+ \frac{p_{\theta_0 j i}}{p_{\theta j i}} \right) - 1 \leq \epsilon_{M_\nu}^2 \right\} \geq \exp(-c_\pi \kappa M_\nu \epsilon_{M_\nu}^2),$$

where $\log_+ x = \max(\log x, 0)$, for $x > 0$.

(A5) (Convexity of metric) The metric, ρ , satisfies that for any positive integer N_ν , $\theta_1, \dots, \theta_N, \theta' \in \Theta$ and non-negative weights, $\sum_{i=1}^{N_\nu} w_i = 1$,

$$\rho \left(\sum_{i=1}^{N_\nu} w_i \theta_i, \theta' \right) \leq \sum_{i=1}^{N_\nu} w_i \rho(\theta_i, \theta').$$

(A6) (Non-zero inclusion probabilities) Define constant $\gamma \geq 1 : \sup_{\nu} \left(\max_{i \in U_{\nu j}} \frac{1}{\pi_{\nu ji}} \right) \leq \gamma$, for all $j = 1, \dots, K$, uniformly, and constants $g_1, g_2 > 0$ where $g_1 \gamma M_{\nu} \leq N_{\nu} \leq g_2 \gamma M_{\nu}$.

(A7) (Asymptotic Independence Condition)

$$\limsup_{\nu \uparrow \infty} \max_{i \neq \ell \in U_{\nu j}} \left| \frac{\pi_{\nu j i \ell}}{\pi_{\nu ji} \pi_{\nu j \ell}} - 1 \right| = O(M_{\nu}^{-1}) \text{ with } P_{\theta_0}\text{-probability 1}$$

such that for some constant, $c_{\nu 3} > 0$, and sufficiently large M_{ν} ,

$$M_{\nu} \sup_{\nu} \max_{i \neq \ell \in U_{\nu j}} \left[\frac{\pi_{\nu j i \ell}}{\pi_{\nu ji} \pi_{\nu j \ell}} - 1 \right] \leq c_{\nu 3}, \text{ for all } j = 1, \dots, K, \text{ uniformly.}$$

(A8) (Constant sampling fraction) For some constant, $f \in (0, 1)$, that we term the “sampling fraction”,

$$\limsup_{\nu} \left| \frac{m_{\nu j}}{M_{\nu j}} - f \right| = O(1), j = 1, \dots, K, \text{ with } P_{\theta_0}\text{-probability 1.}$$

A few comments about our assumptions are in order. Assumptions (A1)–(A5) follow from Srivastava et al. (arXiv:1508.05880). The compactness of Θ in (A1) allows setting $\Theta = \Theta_{M_{\nu}}$, the countable sequence of model spaces. Theorem 4.1 will show the L_1 contraction of the subset pseudo posteriors to the delta measure centered on θ_0 with P_{θ_0} -probability 1. Assumption (A4) imposes a stronger exponential decay control over the tail probability than the condition that averages L_2 norms of the log-likelihood ratio evaluated at the finite population data values specified in Theorem 4 of Ghosal & van der Vaart (2007); however, we still use these assumptions for easy comparisons between the results in this work and in Srivastava et al. (arXiv:1508.05880). There is no loss of generality as the result goes through with the condition from Theorem 4 of Ghosal & van der Vaart (2007) with minor modifications.

Assumptions (A6)–(A8) are the same as those used in Savitsky & Toth (2016) and, together, impose conditions on the sampling distribution, P_{ν} , that define a restricted class of sampling designs. Assumption (A6) requires the sampling design to assign a positive probability for inclusion of every unit in the finite population. No portion of the population may be

systematically excluded, which would prevent a sample of any size from containing information about the population from which the sample is taken. Assumption (A7) restricts the result to sampling designs where the dependence among lowest-level sampled units attenuates to 0 as $\nu \uparrow \infty$; for example, a two-stage sampling design of clusters within strata would meet this condition if the number of population units nested within each cluster from which the sample is drawn increases in the limit of ν . Assumption (A8) ensures that the observed sample size in each subset, m_ν , limits to ∞ along with the size of the partially-observed finite population to which the subset links, M_ν .

Theorem 4.1. *Suppose assumptions (A1)–(A8) hold for subset pseudo posteriors, $\Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]})$ ($j = 1, \dots, K$). Then there exist positive constants $c_{\nu 1}, r_1, r_{\nu 2}, \gamma, c_4$ and large constant, $B_0 = \max[\rho(\theta, \theta_0)]$, $\theta \in \Theta$, such that for sufficiently large M_ν ,*

$$E_{P_{\theta_0}, P_\nu} [W_2^2 \{ \Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0}(\cdot) \}] \leq c_{\nu 1}^2 \epsilon_{M_\nu}^2 + B_0 \left[\frac{1}{r_{\nu 2} M_\nu \epsilon_{M_\nu}^2} + 5 \exp(-r_1 c_4 N_\nu \epsilon_{M_\nu}^2) \right], \quad (8)$$

uniformly for all $j = 1, \dots, K$, where $r_1 \geq \frac{(c_\pi g_2 + 3(\kappa\gamma)^{-1})}{g_1}$, $r_{\nu 2} = \frac{1}{[c_{\nu 3} + \gamma]} \leq 1$, $c_{\nu 1} = \sqrt{\frac{2r_1 g_2 \gamma^2}{q_1 f C_L}}$, $c_4 = \min\left(\frac{q_2}{q_1}, 1\right)$.

We note that the rate of convergence is injured for a sampling distribution, P_ν , that assigns relatively low inclusion probabilities to some units in the finite population such that γ will be relatively larger. Constants r_1 and $r_{\nu 2}$ decrease, while $c_{\nu 1}$ increases as γ becomes larger. Samples drawn under a design that induces a large variability in the sampling weights will express more dispersion in their information similarity to the underlying finite population. Similarly, the larger the dependence among the finite population unit inclusions induced by P_ν , the higher will be $c_{\nu 3}$ and the slower will be the rate of contraction.

Theorem 4.2. *Suppose conditions (A1)–(A8) hold for subset pseudo posteriors, $\Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]})$*

($j = 1, \dots, K$). Then as $M_\nu \uparrow \infty$,

$$W_2 \left\{ \overline{\Pi}^\pi(\cdot \mid \{y_i : \delta_{\nu i} = 1, i = 1, \dots, N_\nu\}), \delta_{\theta_0}(\cdot) \right\} = O_P(\epsilon_{M_\nu}), \quad (9)$$

where O_P is in (P_{θ_0}, P_ν) -probability.

In practice, one may try to plug-in a value for ϵ_{M_ν} that satisfies the conditions, $\epsilon_{M_\nu} \downarrow 0$ and $M_\nu \epsilon_{M_\nu}^2 \uparrow \infty$ and $m_\nu \epsilon_{M_\nu}^2 \uparrow \infty$ as the positive integer $\nu \uparrow \infty$, to the bound in Theorem 4.1 and the convergence order in Theorem 4.2 to see if the resultant bound limits to 0; for example, choosing $\epsilon_{M_\nu} = (\log^2 M_\nu / M_\nu)^{1/2}$, used by Srivastava et al. (arXiv:1508.05880) for so-called regular models; for example, models with continuous densities, which are the class of models we specify in our Assumption (A2) works in both Theorems.

These two theorems extend similar results of Srivastava et al. (arXiv:1508.05880) for independent data to dependent data, where dependence is induced through the sampling design distribution, P_ν ; for example, sampling without replacement and we use Assumptions (A6)–(A8) to accomplish this. The proofs of both theorems generally follows from the techniques in Srivastava et al. (arXiv:1508.05880) with substantial modifications to account for informative sampling and the sampling design-induced dependence among the observations. Our approaches include two unique enabling lemmas. Proofs of the two theorems are in the Appendix and the proofs of enabling lemmas are in Section 1 of the Supplementary Material.

5 Data Analysis

5.1 Hierarchical Model for Current Employment Statistics Survey Data

Our motivating data consists of survey responses in the state of California in a 12 month period from October, 2010 to September, 2011. Let c index an establishment-by-month case observation for establishment i and in month $t\{i\}$ ($i = 1, \dots, n$; $t\{i\} = 1, \dots, T_i$; $n_c = \sum_{i=1}^n T_i$;

$c = 1, \dots, n_c$). Let $T = \max(T_1, \dots, T_n)$ denote the number of unique months observed in the data. Let ℓ ($\ell = 1, \dots, L$) index the number of industries. We define industries using the North American Industry Classification System, which assigns a 6-digit code over 1100 industries. We use the first two digits that denote the industry “super-sectors” for our data. There are $L = 23$ super-sectors populated by $n = 36390$ establishments in California; see Table 1 in the Supplementary Material for the definition of the super-sectors and the allocation of establishments.

The goal for our modeling is to use the temporal- and industry-indexed dependence structures among establishments to efficiently impute missing values for one or more employment count statistics. Noting that the *total number of employees* and the *total number of production workers* reported in the survey are highly dependent, we define a two-dimensional response including the two responses. The fraction of missing responses for the total number of employees in the survey is 45 out of 294674. This is much smaller than the fraction of missing responses for the total number of production workers, which equals 142999 out of 294674. Accounting for the dependence between the total number of employees and the total number of production workers leads to better imputation of missing responses than the case where dependence between the two responses is ignored.

We next construct a negative binomial sampling-weighted *pseudo* likelihood for the observed sample of establishment employment counts from our survey data with

$$y_{cq} \mid \tau_q, \psi_{cq} \stackrel{\text{ind}}{\sim} \text{NB} \{ \tau_q, \exp(\psi_{cq}) \}^{\tilde{w}_{i\{c\}}} , \quad (c = 1, \dots, n_c; q = 1, \dots, Q)$$

$$\psi_c = \theta_{t\{c\}} + \gamma_{\ell\{c\}t\{c\}} z_c, \quad (10)$$

where $\stackrel{\text{ind}}{\sim}$ denotes “independently sampled from,” $Q = 2$, $\tilde{w}_{i\{c\}}$ is the informative sampling weight for establishment i linked to case c , and NB represents the negative binomial distribution with τ_q and $\exp(\psi_{cq})$ as its size and mean parameters. The indexing of precision parameters, τ_q ($q = 1, \dots, Q$), by employment count response variable, q , permits the by-variable

modeling of over-dispersion present in each employment count variable from our data due to the large variation in the sizes of establishments in both the population and sample. The $Q \times 1$ mean on the logarithm scale, ψ_c , is constructed from multivariate fixed and random effects. The subscripts, $t\{c\}$, $\ell\{c\}$, and $i\{c\}$, used to construct the mean on the logarithm scale in (10) denote the month t , industry ℓ , and establishment i linked to case observation c ($t = 1, \dots, T$; $\ell = 1, \dots, L$; $i = 1, \dots, n$; $c = 1, \dots, n_c$). Fixed effect intercepts are denoted by the $Q \times T$ matrix $\Theta = (\theta_1, \dots, \theta_T)$, indexed by response variable and month. We specify industry indexed $Q \times T \times L$ random effects array, $\Gamma = (\Gamma_1, \dots, \Gamma_L)$, where the $Q \times 1$ vector $\gamma_{\ell t}$ models an effect for industry ℓ in month t , ($\ell = 1, \dots, L$; $t = 1, \dots, T$). Random effects predictor, z_c , represents the total employment for establishment, $i\{c\}$, on a 6 month lagged basis in month, $t\{c\}$, obtained from a census instrument, the Quarterly Census of Employment and Wages. The 6 month lag derives from the relatively rapid Current Employment Statistics production schedule under which employment statistics are published on a more timely basis than for the Quarterly Census of Employment and Wages. The historical values the Quarterly Census of Employment and Wages serve as a magnitude variable. The two terms of (10) allow for non-linear associations over industries and months to each response variable.

We complete the specification of our probability model with the following priors,

$$\Theta^{Q \times T} \sim \mathcal{N}_{Q \times T}(0, P_2^{-1} \circ P_3^{-1}), \quad \Gamma_\ell^{Q \times T} \stackrel{\text{iid}}{\sim} \mathcal{N}_{Q \times T}(0, P_8^{-1} \circ P_6^{-1}) \quad (\ell = 1, \dots, L), \quad (11a)$$

$$P_s \sim \text{Huang-Wand}(\nu, b_{s1}, \dots, b_{sQ}), \quad b_{sq} \stackrel{\text{iid}}{\sim} \mathcal{G}(1/2, 1), \quad q = 1, \dots, Q \quad (s = 2, 8), \quad (11b)$$

$$P_s = D - \rho_s \Omega; \quad \rho_s \sim \mathcal{U}(0, 1) \quad (s = 3, 6), \quad \tau_q^{-1/2} \stackrel{\text{iid}}{\sim} \mathcal{C}(0, 1), \quad (11c)$$

where $\stackrel{\text{iid}}{\sim}$ denotes “independently and identically distributed as,” Huang-Wand is a marginally noninformative prior for covariance matrices (Huang & Wand 2013), and \mathcal{N} , \mathcal{U} , and \mathcal{C} denote the Gaussian, uniform, and Cauchy distributions. The matrix Ω is a $T \times T$ adjacency matrix where $\omega_{ij} = 1$ if months i and j are adjacent; else, $\omega_{ij} = 0$, and D is a $T \times T$ diagonal matrix of row sums of Ω such that the precisions for months with a larger number of neigh-

bors will be higher than those with a relatively smaller number of neighbors. The form of the priors and the algorithm to sample from the pseudo posterior distribution of parameters $\{\Theta, \Gamma_1, \dots, \Gamma_L, \tau_1, \dots, \tau_Q\}$ are described in Section 2 of the Supplementary Material.

5.2 Setup and Comparison Metric

We compared the performance of our method with the full-sample pseudo posterior distribution. The sampling model for the simulated and real data were based on the hierarchical model in (10). The sampling algorithm described in Section 2 of the Supplementary Material was used to obtain samples from every posterior distribution after appropriately choosing the sampling weights w_{ij} in (10). All sampling algorithms ran for 15,000 iterations. We collected every fifth sample after discarding the first 10,000 samples as burn-ins. The convergence of every chain to its stationary distribution was confirmed using trace plots. Sampling from the subset posterior and full-data pseudo posterior distributions respectively required 8GB and 32GB of memory resources.

We used a metric based on the total variation distance to evaluate the accuracy of approximation of a probability distribution P by another probability distribution Q . Let p and q be the densities of P and Q with respect to the Lebesgue measure, then the accuracy of Q in approximating P was defined as

$$\text{accuracy}(q) = 1 - \frac{1}{2} \int_{-\infty}^{\infty} |p(y) - q(y)| dy. \quad (12)$$

The $\text{accuracy}(q)$ metric belonged to the interval $[0, 1]$ (Faes et al. 2012) and was computed using numerical approximation. The approximation of P by Q was excellent or poor if $\text{accuracy}(q)$ was close to 1 or 0, respectively.

5.3 Simulated Data

Consider the sampling model of the Current Employment Statistics survey data described in Section 5.1. We fixed N, T, Q , and L defined in Section 5.1 as 10,000, 10, 2, and 1, which excluded any industry-indexed random effects without loss of generality. We fixed ρ at 0.9 to simulate P_3 using (11c). Given t , we fixed $\text{var}(y_{it1})$, $\text{var}(y_{it2})$, and $\text{cov}(y_{it1}, y_{itq})$ at 0.5, 2, and 0.6 to define P_2 ($i = 1, \dots, N$). We first simulated Θ using (11a) and then generated the population level response q for establishment i at time t , y_{itq} ($i = 1, \dots, N; q = 1, \dots, Q$), as follows:

$$y_{itq} \mid \tau_q, \psi_{tq} \stackrel{\text{ind}}{\sim} \text{NB} \{ \tau_q, \exp(\psi_{tq}) \}, \quad \psi_{tq} = 5 + \theta_{tq}, \quad (13)$$

where $i = 1, \dots, 10,000$, $t = 1, \dots, 10$, $q = 1, 2$, $\tau_1 = 5$, and $\tau_2 = 10$. The covariance matrices P_2 and P_3 induced dependence in θ_{tq} s across ts and the two qs .

We first generated a finite population according to (13), then subsequently drew two informative samples from the finite population of the N establishments with the inclusion probability for each establishment i set to be proportional to $y_{i..} = \sum_{t=1}^T \sum_{q=1}^Q y_{itq}$. The sampled data are composed of response values for both variables and all 10 time points for each establishment included in each sample. We sampled $n = fN$ of the N establishments of the finite population in each of the two samples for sampling fraction, $f \in \{0.4, 0.6\}$. Establishments contained in each of the two samples were next randomly partitioned into K subsets, each of equal size, $m = n/K$, where $K \in \{5, 10\}$.

We next obtained samples of parameters under (10) from the finite population posterior distribution, full-sample pseudo posterior distribution, and our method in every replication. We set $w_i = 1$ to obtain parameter draws from the finite population posterior distribution. Parameter draws from the full-sample pseudo posterior distribution of size n were estimated by setting $w_i = n y_{i..}^{-1} (\sum_{i=1}^n y_{i..}^{-1})^{-1}$ ($i = 1, \dots, n$), which normalizes the sampling weights to sum to n for regulation of the uncertainties of estimated parameters. We drew parameter

samples from subset pseudo posterior j by normalizing $w_{ij} = n y_{i..}^{-1} (\sum_{i=1}^n y_{i..}^{-1})^{-1}$ for every establishment i in the j th subset in (10), which regulates the amount of uncertainty in each subset j to approximate that in the full sample. Next, the samples from all subset posterior distributions were combined using the algorithm in Li et al. (arXiv:1605.04029). This simulation setup was replicated 10 times.

The generalized Wasserstein posterior showed excellent performance in approximating the full-data pseudo posterior distribution for both $K = 5$ and $K = 10$. Figure 1 demonstrates that estimated pseudo posterior densities our method under both $K = 5$ and $K = 10$ very closely approximate the full-sample pseudo posterior, both in locations and the amount of estimated uncertainties. The full data and Table 1 displays computed accuracies for the θ_{qt} s, which are all close to 1. Assumptions (A1)–(A8) were satisfied in our simulation example, so the results of our method were not sensitive to the size of the subsets K , agreeing with Theorem 4.2. The conditions of Theorem 4.1 were easier to satisfy when $K = 5$ than when $K = 10$ due to a larger subset size for $K = 5$, resulting in higher accuracy for the generalized Wasserstein posterior with $K = 5$ in certain cases. In all our simulation examples, the generalized Wasserstein posterior required only 25% of the memory resources used by the full-data pseudo posterior and was about 10-times faster than the full-data pseudo posterior in run-time (Figure 2).

5.4 Application to Current Employment Statistics Survey Data

The survey data for California had $n = 39360$ business establishments, each providing responses over multiple months for a total of $n_c = 297000$ establishment-month cases. We used the establishment-month case observations in the state of California for our comparisons because it was computationally feasible to estimate the full-sample pseudo posterior distribution using the hierarchical model in (10). Our goal was to demonstrate that the generalized Wasserstein barycenter could be used as an alternative for the full-data pseudo posterior distribution

for inference on model parameters and for imputation of missing responses.

We randomly allocated the n establishments to $K = 4$ subsets of roughly equal numbers of establishments, $m = (9017, 9140, 9082, 9151)$ associated with $n_c = (72841, 74009, 73702, 74122)$ establishment-month case observations. We divided the n establishments into $L = 23$ industry-indexed strata and conducted simple random sampling within each stratum to populate the subsets. Stratified selection ensured that all $L = 23$ industry super-sectors were linked to one or more establishments in each subset. We selected 4 subsets to accommodate our budget for computation and to ensure that m_j was sufficiently large such that the conditions for our Theorem 4.1 were satisfied.

The generalized Wasserstein posterior provided an excellent approximation to the full-data pseudo posterior distribution. The marginals of the generalized Wasserstein posterior and the full-data pseudo posterior distribution were fairly similar across various industry super-sectors (Figure 3). The scaling of the subset pseudo posteriors under generalized stochastic approximation worked very well in that the spread of generalized Wasserstein posterior and full-data pseudo posterior distributions were very similar, suggesting that uncertainty quantification using the two posterior distributions would be very similar. This was further confirmed using the metric in (12), which showed that the generalized Wasserstein posterior was more than 90% accurate in approximating the marginals of the full-data pseudo posterior for Θ (Table 2).

The generalized Wasserstein posterior also showed excellent performance in imputation. Our model in (10) involved specification of a relatively large number of parameters to parameterize the log means, ψ_{cq} s. We constructed the means of our negative binomial model on the data scale, $\exp(\psi_{cq})$, using (10). These means were used to impute the missing y_{cq} s from the posterior predictive distribution constructed from the generalized Wasserstein posterior for θ_{qj} s and $\gamma_{\ell qj}$ s. The distribution of the posterior mean values of $\exp(\psi_{cq})$ s associated with the missing responses were nearly identical for the full-data pseudo posterior distribution and the generalized Wasserstein posterior (Figure 4).

6 Concluding Remarks

The efficiency of the generalized Wasserstein posterior was critical in extending the inference on a low-dimensional parameter space to imputation on a parameter space of medium dimensions that provided sufficient flexibility for high quality imputation. Future areas of exploration include assessing feasibility of generalized Wasserstein posterior under joint modeling of marginal sampling weights and the response of interest in a fully Bayesian construction, as contrasted with the plug-in pseudo posterior.

A Proof of Theorem 4.1

We begin the proof in the same manner as in Srivastava et al. (arXiv:1508.05880) by deconstructing the expectation of the squared Wasserstein distance from the pseudo posterior for subset j , $\Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]})$, to the delta measure at θ_0 , into two parts. We recall assumption (A1) that Θ is compact, so that the sieve, Θ_{N_ν} , specified in Ghosal & van der Vaart (2007) equals the entire space, Θ , and we are able to bound, $\rho(\theta, \theta_0) < B_0$:

$$\begin{aligned}
E_{P_{\theta_0}, P_\nu} [W_2^2 \{ \Pi_j^\pi(\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0}(\cdot) \}] &= E_{P_{\theta_0}, P_\nu} \int_{\theta \in \Theta} \rho^2(\theta, \theta_0) \Pi_j^\pi(d\theta \mid y_{[j]} \delta_{\nu[j]}) \\
&\leq E_{P_{\theta_0}, P_\nu} \int_{\{\theta: \rho(\theta, \theta_0) \leq c_{\nu 1} \epsilon_{M_\nu}\}} \rho^2(\theta, \theta_0) \Pi_j^\pi(d\theta \mid y_{[j]} \delta_{\nu[j]}) + E_{P_{\theta_0}, P_\nu} \int_{\{\rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu}\}} \rho^2(\theta, \theta_0) \Pi_j^\pi(d\theta \mid y_{[j]} \delta_{\nu[j]}) \\
&\leq (c_{\nu 1} \epsilon_{M_\nu})^2 + B_0^2 E_{P_{\theta_0}, P_\nu} \Pi_j^\pi(\rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]}). \tag{14}
\end{aligned}$$

We set constant, $c_{\nu 1} = \sqrt{\frac{2r_1 g_2 \gamma^2}{q_1 f C_L}}$, similar to Srivastava et al. (arXiv:1508.05880), though we note that it now depends on the upper bound on the sampling weights, γ , specified in assumption (A6) over all $(i, j) \in U_{\nu j}$ ($j = 1, \dots, k$). Constant $c_{\nu 1}$ also depends on the limiting sampling fraction, f , for each subset; see assumption (A8). The additional constants g_2, r_1, q_1, C_L are specified in assumptions (A2) and (A6) and in Lemmas 1 and 2 in the Sup-

plementary Material.

We next focus to bound the second term on the right-hand side of (14). The flow of the proof is most similar to Theorem 4.3 of Srivastava et al. (arXiv:1508.05880) and Theorem 3 of Savitsky & Toth (2016). We extend these approaches to account for the taking of an informative random sample from the finite sub-populations, $U_{\nu j}$ ($j = 1, \dots, K$). We first use assumption (A2) to bound the pseudo posterior with respect distance metric ρ from above by the pseudo Hellinger distance,

$$\begin{aligned} \Pi_j^\pi \left(\theta \in \Theta : \rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right) &\leq \Pi_j^\pi \left(\theta \in \Theta : h_{M_{\nu j}}(\theta, \theta_0) > \sqrt{C_L} c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right) \\ &\leq \Pi_j^\pi \left(\theta \in \Theta : h_{M_{\nu j}}^\pi(\theta, \theta_0) > \sqrt{C_L} c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right), \end{aligned} \quad (15)$$

where last inequality results because $h_{M_{\nu j}}^\pi \geq h_{M_{\nu j}}$.

We next bound the expectation with respect to the joint distribution, (P_{θ_0}, P_ν) , of the pseudo posterior,

$$\begin{aligned} \Pi_j^\pi \left(\theta \in \Theta : h_{M_{\nu j}}^\pi(\theta, \theta_0) > \sqrt{C_L} c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right) &= \\ \frac{\int_{\{\theta \in \Theta : h_{M_\nu}^\pi \geq \sqrt{C_L} c_{\nu 1} \epsilon_{M_\nu}\}} \prod_{i=1}^{M_\nu} \frac{p_{\theta j i}^\pi}{p_{\theta_0 j i}^\pi} \Pi(d\theta)}{\int_{\theta \in \Theta} \prod_{i=1}^{M_\nu} \frac{p_{\theta j i}^\pi}{p_{\theta_0 j i}^\pi} \Pi(d\theta)}. \end{aligned} \quad (16)$$

We may bound the probability mass from below for some minimum value of the denominator of (16) using assumption (A4) and Lemma 2 in the Supplementary Material such that with probability greater than or equal to $1 - (r_{\nu 2} M_\nu \epsilon_{M_\nu}^2)^{-1}$,

$$\int_{\theta \in \Theta} \prod_{i=1}^{M_\nu} \frac{p_{\theta j i}^\pi}{p_{\theta_0 j i}^\pi} \Pi(d\theta) > \exp(-r_1 N_\nu \epsilon_{M_\nu}^2). \quad (17)$$

We next bound the numerator of (16), from above, in (P_{θ_0}, P_ν) -probability, using assump-

tions (A3), (A6), and Lemma 1 in the Supplementary Material where the numerator,

$$\int_{\{\theta \in \Theta : h_{M_\nu}^\pi \geq \sqrt{C_L} c_{\nu 1} \epsilon_{M_\nu}\}} \prod_{i=1}^{M_\nu} \frac{p_{\theta j i}^\pi}{p_{\theta_0 j i}^\pi} \Pi(d\theta) \leq \exp\left(-\frac{q_1 m_\nu C_L c_{\nu 1}^2 \epsilon_{M_\nu}^2}{\gamma}\right) \quad (18)$$

$$\leq \exp\left(-2r_1 g_2 \gamma M_\nu \epsilon_{M_\nu}^2\right) \quad (19)$$

$$\leq \exp\left(-2r_1 N_\nu \epsilon_{M_\nu}^2\right). \quad (20)$$

The inequality in (18) results from plugging in $\tau = \sqrt{C_L} c_{\nu 1} \epsilon_{M_\nu}$ into the result for Lemma 1 in the Supplementary Materials. The inequality in (19) results from plugging in for $c_{\nu 1}$. We used $N_\nu \leq g_2 \gamma M_\nu$ from assumption (A6) to achieve the inequality in (20).

The lower bound of (20) is realized with probability at least

$$1 - 4 \exp\left(-\frac{q_2 m_\nu C_L c_{\nu 1}^2 \epsilon_{M_\nu}^2}{\gamma}\right) \quad (21)$$

$$= 1 - 4 \exp\left(-\frac{r_1 q_2 g_2 \gamma M_\nu \epsilon_{M_\nu}^2}{q_1}\right), \quad (22)$$

where we, again, plug in for τ for the probability bound of Lemma 1 in the Supplementary Material to achieve (21) and for $c_{\nu 1}$ to achieve (22). Then with probability at least

$$1 - 4 \exp\left(-\frac{r_1 q_2 g_2 \gamma M_\nu \epsilon_{M_\nu}^2}{q_1}\right) - (r_{\nu 2} M_\nu \epsilon_{M_\nu}^2)^{-1}$$

$$\begin{aligned} & \Pi_j^\pi(\theta \in \Theta : \rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]}) \\ & \leq \exp(-2r_1 N_\nu \epsilon_{M_\nu}^2) \times \exp(r_1 N_\nu \epsilon_{M_\nu}^2) \\ & \leq \exp(-r_1 N_\nu \epsilon_{M_\nu}^2) \end{aligned} \quad (23)$$

Let the event, $A_{M_\nu}^\pi = \{\Pi_j^\pi(\theta \in \Theta : \rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]}) \leq \exp(-r_1 N_\nu \epsilon_{M_\nu}^2)\}$,

which we use to establish the L_1 bound,

$$\begin{aligned}
& E_{P_{\theta_0}, P_\nu} \left[\Pi_j^\pi \left(\theta \in \Theta : \rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right) \right] \\
& E_{P_{\theta_0}, P_\nu} \left[\mathbf{I} \left(A_{M_\nu}^\pi \right) \times \Pi_j^\pi \left(\theta \in \Theta : \rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right) \right. \\
& \quad \left. + \mathbf{I} \left([A_{M_\nu}^\pi]^c \right) \times \Pi_j^\pi \left(\theta \in \Theta : \rho(\theta, \theta_0) > c_{\nu 1} \epsilon_{M_\nu} \mid y_{[j]} \delta_{\nu[j]} \right) \right] \\
& \leq \exp \left(-r_1 N_\nu \epsilon_{M_\nu}^2 \right) + P_{P_{\theta_0}, P_\nu} \left([A_{M_\nu}^\pi]^c \right) \\
& \leq \exp \left(-r_1 N_\nu \epsilon_{M_\nu}^2 \right) + 4 \exp \left(-\frac{r_1 q_2 g_2 \gamma M_\nu \epsilon_{M_\nu}^2}{q_1} \right) + \frac{1}{r_{\nu 2} M_\nu \epsilon_{M_\nu}^2} \\
& \leq \exp \left(-r_1 N_\nu \epsilon_{M_\nu}^2 \right) + 4 \exp \left(-\frac{r_1 q_2 N_\nu \epsilon_{M_\nu}^2}{q_1} \right) + \frac{1}{r_{\nu 2} M_\nu \epsilon_{M_\nu}^2} \tag{24}
\end{aligned}$$

$$\leq 5 \exp \left(-r_1 c_4 N_\nu \epsilon_{M_\nu}^2 \right) + \frac{1}{(r_{\nu 2} M_\nu \epsilon_{M_\nu}^2)}, \tag{25}$$

where $c_4 = \min \left(\frac{q_2}{q_1}, 1 \right)$. The first term in (25) dominates because our loss of independence prevents the use of Bernstein's inequality as leveraged in Massart (2007) and Srivastava et al. (arXiv:1508.05880) to get an exponential lower bound on the denominator of (16). So we are not able to demonstrate an optimal rate of convergence. Returning to the decomposition of the W_2 distance in (14),

$$\begin{aligned}
& E_{P_{\theta_0}, P_\nu} W_2^2 \left(\Pi_j^\pi \left(\cdot \mid y_{[j]}, \delta_{\nu[j]} \right), \delta_{\theta_0}(\cdot) \right) \\
& \leq (c_{\nu 1} \epsilon_{M_\nu})^2 + B_0 \left[\frac{1}{r_{\nu 2} M_\nu \epsilon_{M_\nu}^2} + 5 \exp \left(-r_1 c_4 N_\nu \epsilon_{M_\nu}^2 \right) \right], \tag{26}
\end{aligned}$$

uniformly for all $j = 1, \dots, K$, for constants, $r_1 \geq \frac{(c_\pi g_2 + 3(\kappa\gamma)^{-1})}{g_1}$, $r_{\nu 2} = \frac{1}{[c_{\nu 3} + \gamma]} \leq 1$, $c_{\nu 1} = \sqrt{\frac{2r_1 q_2 \gamma^2}{q_1 f C_L}}$ and $c_4 = \min \left(\frac{q_2}{q_1}, 1 \right)$.

B Proof of Theorem 4.2

Proof. We bound the probability using Chebyshev, and Lemma B.7 from Srivastava et al. (arXiv:1508.05880), such that for any constant, $c_{\nu 5}$ (a function of the sampling design con-

stants, $(\gamma, c_{\nu 3})$,

$$P_{\theta_0, P_\nu} (W_2 (\bar{\Pi}^\pi (\cdot \mid \{y_i : \delta_{\nu i} = 1, i = 1, \dots, N_\nu\}), \delta_{\theta_0} (\cdot)) > \sqrt{c_{\nu 5} \epsilon_{M_\nu}}) \quad (27a)$$

$$\leq P_{\theta_0, P_\nu} \left(\frac{1}{K} \sum_{j=1}^K W_2 (\Pi_j^\pi (\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0} (\cdot)) > \sqrt{c_{\nu 5} \epsilon_{M_\nu}} \right) \quad (27b)$$

$$\leq \frac{1}{c_{\nu 5} \epsilon_{M_\nu}^2} \text{var}_{\theta_0, P_\nu} \left[\frac{1}{K} \sum_{j=1}^K W_2 (\Pi_j^\pi (\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0} (\cdot)) \right] \quad (27c)$$

$$\leq \frac{1}{c_{\nu 5} \epsilon_{M_\nu}^2} E_{\theta_0, P_\nu} \left[\left(\frac{1}{K} \sum_{j=1}^K W_2 (\Pi_j^\pi (\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0} (\cdot)) \right)^2 \right] \quad (27d)$$

$$= \frac{1}{c_{\nu 5} \epsilon_{M_\nu}^2 K^2} \left\{ \sum_{\ell=j=1}^K E_{\theta_0, P_\nu} W_2^2 (\Pi_j^\pi (\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0} (\cdot)) \right. \quad (27e)$$

$$\left. + \sum_{\ell \neq j=1}^K E_{\theta_0, P_\nu} [W_2 (\Pi_\ell^\pi (\cdot \mid Y_{[\ell]} \delta_{\nu \ell}), \delta_{\theta_0} (\cdot)) \cdot W_2 (\Pi_j^\pi (\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0} (\cdot))] \right\} \quad (27f)$$

$$\leq \frac{1}{c_{\nu 5} \epsilon_{M_\nu}^2 K^2} \cdot K^2 E_{\theta_0, P_\nu} \max_{j \in (1, \dots, K)} W_2^2 (\Pi_j^\pi (\cdot \mid y_{[j]}, \delta_{\nu[j]}), \delta_{\theta_0} (\cdot)) \quad (27g)$$

$$\leq \frac{1}{c_{\nu 5} \epsilon_{M_\nu}^2} \cdot \left[c_{\nu 1}^2 \epsilon_{M_\nu}^2 + B_0 \left(\frac{1}{r_{\nu 2} M_\nu \epsilon_{M_\nu}^2} + 5 \exp(-r_1 c_4 N_\nu \epsilon_{M_\nu}^2) \right) \right] \quad (27h)$$

$$\leq \frac{3B_0 c_{\nu 1} \epsilon_{M_\nu}^2}{c_{\nu 5} \epsilon_{M_\nu}^2} = \frac{3B_0 c_{\nu 1}}{c_{\nu 5}}, \quad (27i)$$

for ν sufficiently large. □

References

Agueh, M. & Carlier, G. (2011), ‘Barycenters in the Wasserstein space’, *SIAM Journal on Mathematical Analysis* **43**(2), 904–924.

Breslow, N. E. & Wellner, J. A. (2007), ‘Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression’, *Scandinavian Journal of Statistics* **34**(1), 86–102.

URL: <http://EconPapers.repec.org/RePEc:bla:scjsta:v:34:y:2007:i:1:p:86-102>

- Dong, Q., Elliott, M. R. & Raghunathan, T. E. (2014), ‘A nonparametric method to generate synthetic populations to adjust for complex sampling design features’, *Survey Methodology* **40**(1), 29–46.
- Faes, C., Ormerod, J. T. & Wand, M. P. (2012), ‘Variational bayesian inference for parametric and nonparametric regression with missing data’, *Journal of the American Statistical Association* **106**(495), 959–971.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013), *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*, third edn, Chapman and Hall/CRC, London.
URL: <http://www.stat.columbia.edu/gelman/arm/missing.pdf>
- Ghosal, S., Ghosh, J. K. & Vaart, A. W. V. D. (2000), ‘Convergence rates of posterior distributions’, *Ann. Statist* pp. 500–531.
- Ghosal, S. & van der Vaart, A. (2007), ‘Convergence rates of posterior distributions for noniid observations’, *Ann. Statist.* **35**(1), 192–223.
URL: <http://dx.doi.org/10.1214/009053606000001172>
- Huang, A. & Wand, M. P. (2013), ‘Simple marginally noninformative prior distributions for covariance matrices’, *Bayesian Anal.* **8**(2), 439–452.
URL: <http://dx.doi.org/10.1214/13-BA815>
- Kunihama, T., Herring, A., Halpern, C. T. & Dunson, D. (2016), ‘Nonparametric bayes modeling with sample survey weights’, *Statistics & Probability Letters* **113**, 41–48.
- Massart, P. (2007), *Concentration inequalities and model selection: Ecole d’Eté de Probabilités de Saint-Flour XXXIII - 2003*, number 1896 in ‘Ecole d’Eté de Probabilités de

Saint-Flour’, Springer-Verlag.

URL: https://books.google.com/books?id=O_gZAQAIAAJ

Rao, J. N. K. & Wu, C. (2010), ‘Bayesian pseudo-empirical-likelihood intervals for complex surveys’, *Journal of the Royal Statistical Society Series B* **72**(4), 533–544.

URL: <http://EconPapers.repec.org/RePEc:bla:jorssb:v:72:y:2010:i:4:p:533-544>

Savitsky, T. D. & Dalal, S. R. (2013), ‘Bayesian non-parametric analysis of multirater ordinal data, with application to prioritizing research goals for prevention of suicide’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63**(4), 539–557.

URL: <http://dx.doi.org/10.1111/rssc.12049>

Savitsky, T. D. & Toth, D. (2016), ‘Bayesian Estimation Under Informative Sampling’, *Electronic Journal of Statistics* **10**(1), 1677–1708.

Si, Y., Pillai, N. S. & Gelman, A. (2015), ‘Bayesian nonparametric weighted sampling inference’, *Bayesian Anal.* **10**(3), 605–625.

URL: <http://dx.doi.org/10.1214/14-BA924>

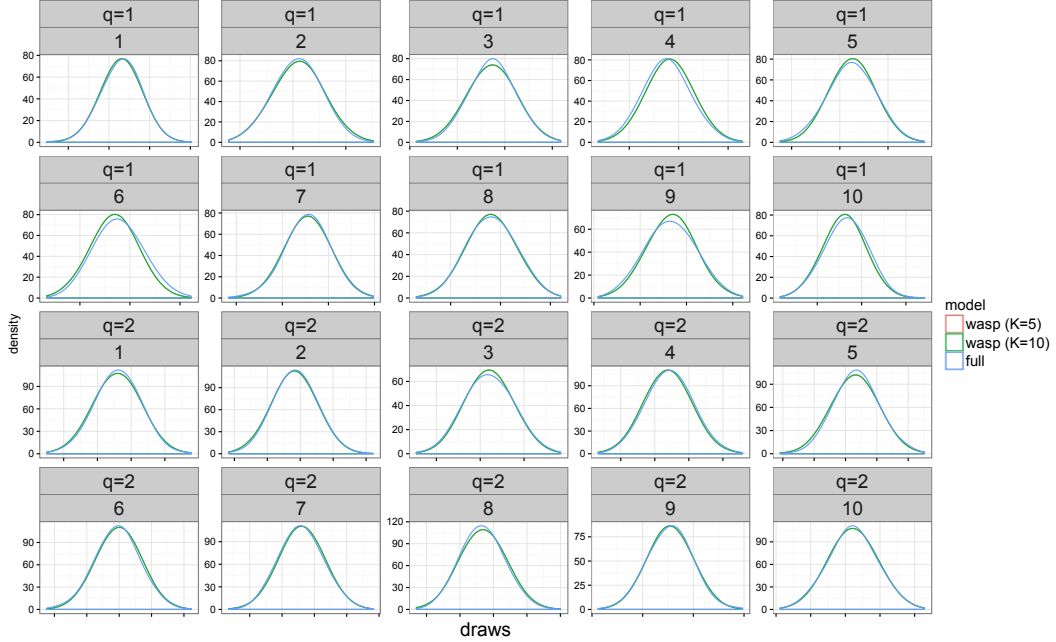
Srivastava, S., Cevher, V., Dinh, Q. & Dunson, D. (2015), WASP: Scalable Bayes via barycenters of subset posteriors, in ‘Proceedings of the 18th International Conference on Artificial Intelligence and Statistics’, pp. 912–920.

Table 1: *The accuracy (12) of the generalized Wasserstein posterior for the marginals of Θ averaged across 10 simulation replications. The maximum Monte Carlo error over 10 simulation replications was 0.045.*

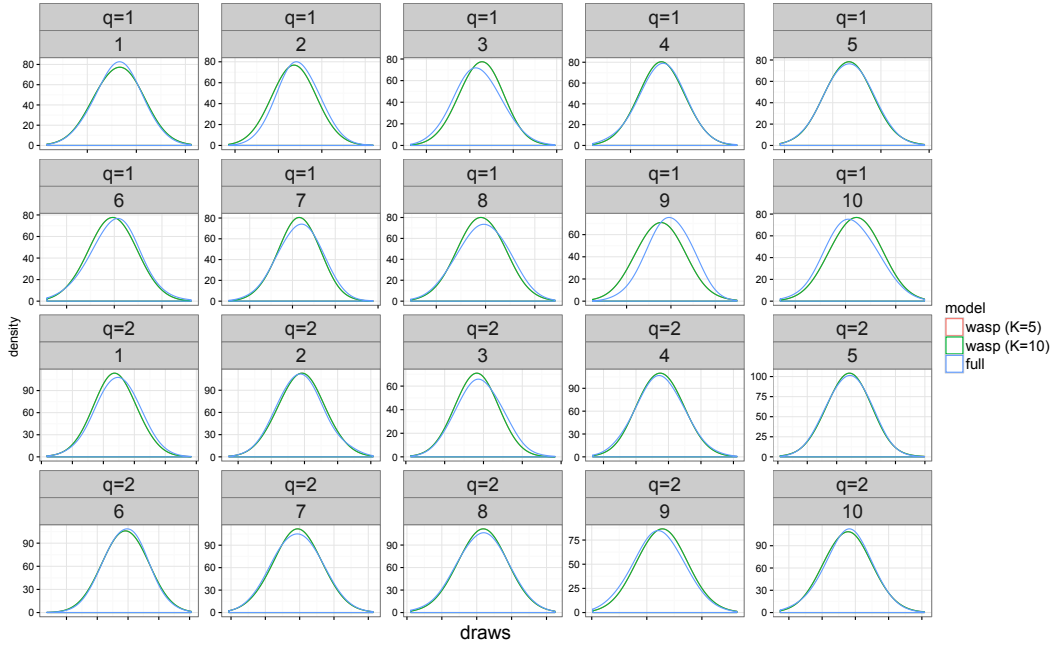
(q, t)	(1, 1)	(2, 1)	(1, 2)	(2, 2)	(1, 3)	(2, 3)	(1, 4)	(2, 4)	(1, 5)	(2, 5)
$K = 5, f = 60\%$	0.96	0.97	0.94	0.97	0.95	0.96	0.95	0.97	0.96	0.96
$K = 10, f = 60\%$	0.95	0.96	0.94	0.96	0.94	0.96	0.95	0.96	0.94	0.96
$K = 5, f = 40\%$	0.96	0.96	0.93	0.96	0.93	0.96	0.95	0.96	0.94	0.96
$K = 10, f = 40\%$	0.95	0.96	0.94	0.95	0.93	0.95	0.95	0.96	0.94	0.96
(q, t)	(1, 6)	(2, 6)	(1, 7)	(2, 7)	(1, 8)	(2, 8)	(1, 9)	(2, 9)	(1, 10)	(2, 10)
$K = 5, f = 60\%$	0.95	0.96	0.96	0.96	0.95	0.96	0.95	0.96	0.96	0.96
$K = 10, f = 60\%$	0.94	0.95	0.95	0.97	0.95	0.96	0.96	0.96	0.96	0.96
$K = 5, f = 40\%$	0.95	0.97	0.94	0.96	0.94	0.96	0.93	0.95	0.95	0.97
$K = 10, f = 40\%$	0.96	0.95	0.93	0.95	0.93	0.95	0.92	0.97	0.95	0.96

Table 2: *The accuracy (12) of the generalized Wasserstein posterior for the marginals θ_{qt} ($q = 1, 2; t = 1, \dots, 24$).*

θ_{11}	θ_{21}	θ_{12}	θ_{22}	θ_{13}	θ_{23}	θ_{14}	θ_{24}	θ_{15}	θ_{25}	θ_{16}	θ_{26}
0.95	0.94	0.98	0.93	0.98	0.93	0.98	0.97	0.98	0.97	0.98	0.97
θ_{17}	θ_{27}	θ_{18}	θ_{28}	θ_{19}	θ_{29}	θ_{110}	θ_{210}	θ_{111}	θ_{211}	θ_{112}	θ_{212}
0.98	0.98	0.96	0.94	0.97	0.97	0.96	0.94	0.98	0.96	0.98	0.94
θ_{113}	θ_{213}	θ_{114}	θ_{214}	θ_{115}	θ_{215}	θ_{116}	θ_{216}	θ_{117}	θ_{217}	θ_{118}	θ_{218}
0.95	0.95	0.98	0.93	0.99	0.93	0.97	0.98	0.97	0.97	0.98	0.97
θ_{119}	θ_{219}	θ_{120}	θ_{220}	θ_{121}	θ_{221}	θ_{122}	θ_{222}	θ_{123}	θ_{223}	θ_{124}	θ_{224}
0.98	0.98	0.97	0.94	0.98	0.96	0.96	0.95	0.98	0.96	0.98	0.95



(a) $f = 60\%$



(b) $f = 40\%$

Figure 1: Comparison of the full-data pseudo posterior density and generalized Wasserstein posterior density with $K = 5$ and $K = 10$ in a simulation replication; full, full-sample pseudo posterior distribution; wasp, generalized Wasserstein posterior distribution; f , the sampling fraction as in assumption (A8); K , the number of subsets.

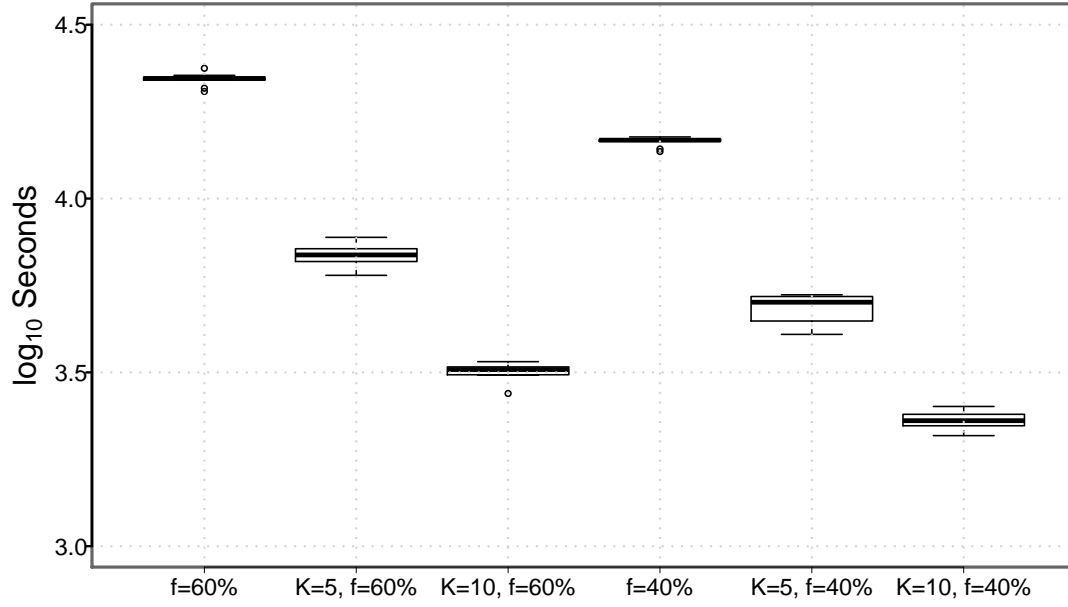
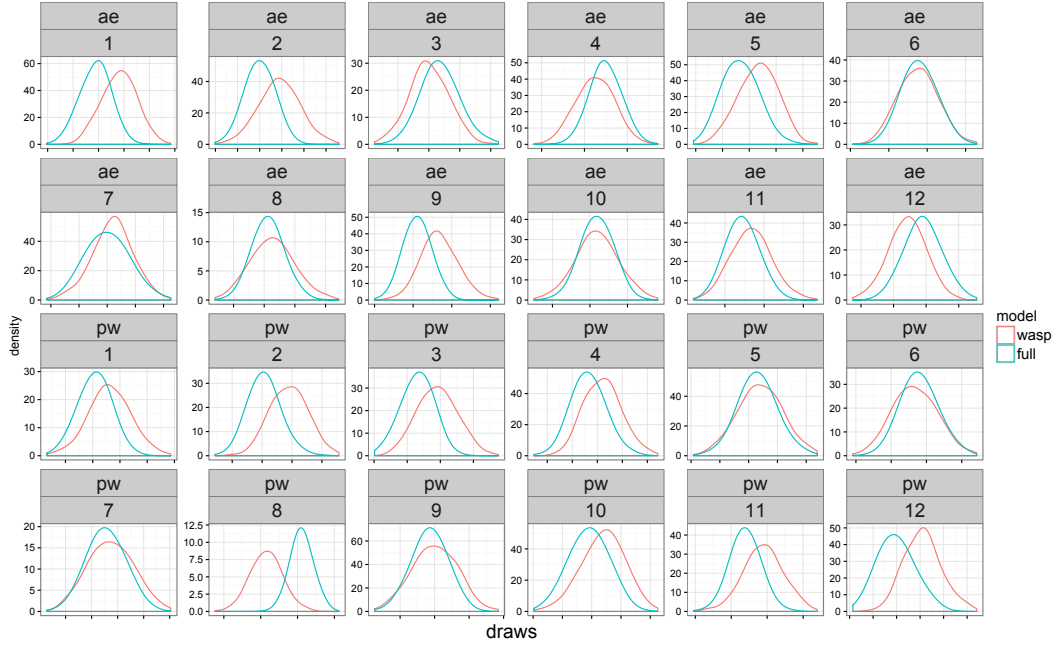
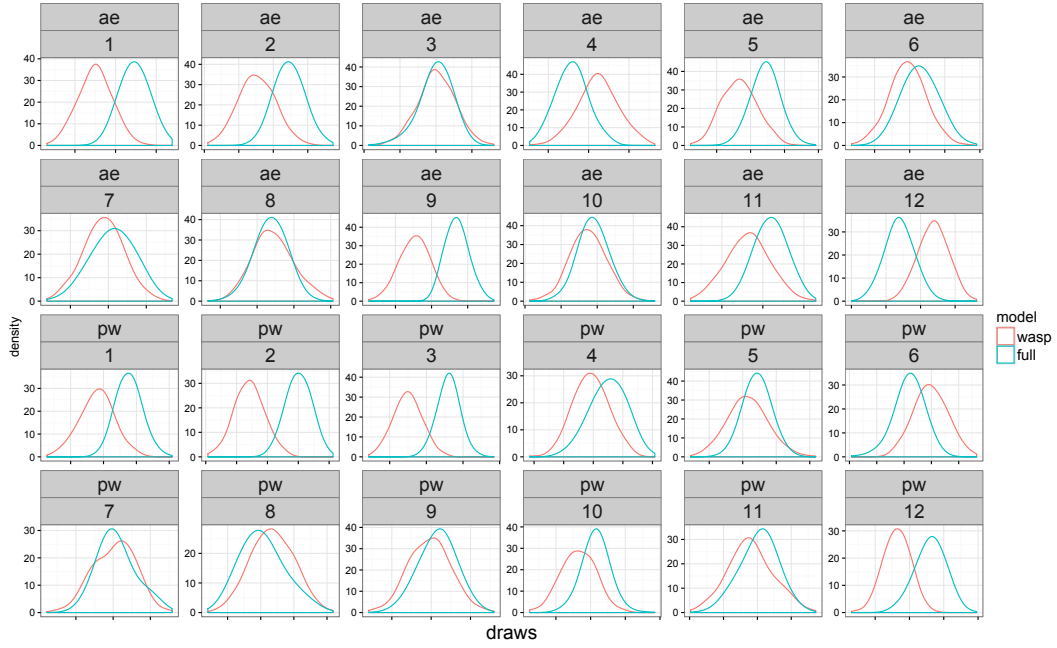


Figure 2: Computation time for the full-data pseudo posteriors and generalized Wasserstein posterior across 10 simulation replications. The x-axis labels with K correspond to generalized Wasserstein posteriors and those without K correspond to full-sample pseudo posteriors; f , the sampling fraction as in assumption (A8); K , the number of subsets.



(a) Γ



(b) Θ

Figure 3: Comparison of generalized Wasserstein posterior and full-data pseudo posterior distribution. Each plot panel compares the generalized Wasserstein posterior of 4 subset pseudo posterior distributions (in red) to the full-data pseudo posterior distribution (in turquoise) for selected parameters. The panels in the top plot are for all $\{\gamma_{\ell qj}\}$, ℓ = the Professional & Technical industry super-sector. The panels in the bottom plot panel include the intercept parameters, $\{\theta_{qj}\}$; full, full-data pseudo posterior distribution; wasp, Generalized Wasserstein

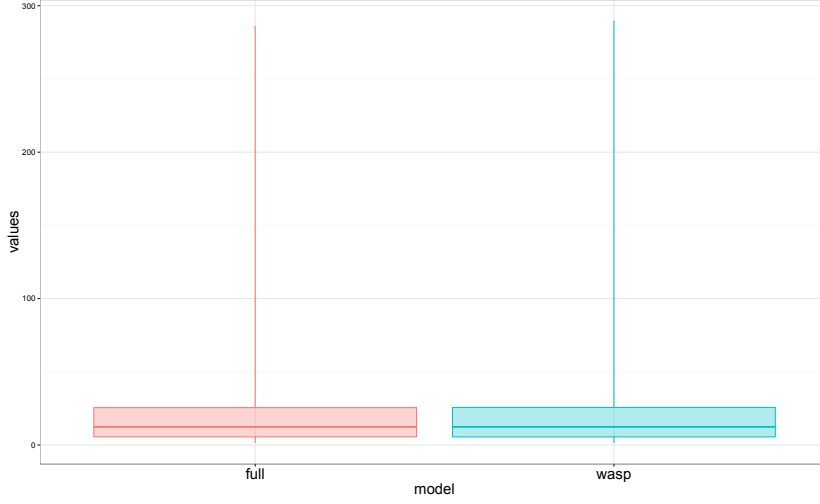


Figure 4: Comparison of distribution of $n_{\text{miss}} \times 1$ posterior means, $\exp(\psi_{cq})$, estimated using the generalized Wasserstein and full-data pseudo posterior distributions. The right-hand boxplot presents the distribution of the $n_{\text{miss}}, \{\exp(\psi_{cq})\}$, linked to missing $\{y_{cq}\}$ estimated from the generalized Wasserstein posterior, while the left-hand boxplot presents the distribution of the $n_{\text{miss}}, \{\exp(\psi_{cq})\}$ estimated from the full-sample pseudo posterior distribution. Approximately 25% of $\{y_{cq}\}$ are missing; full, full-data pseudo posterior distribution; wasp, generalized Wasserstein posterior.